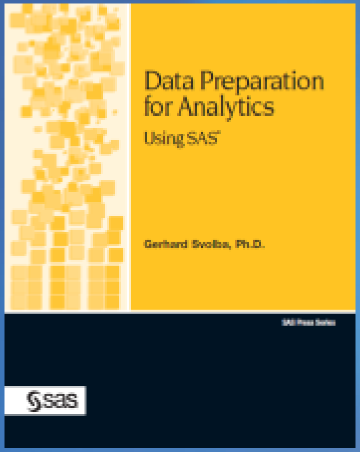
Analytics2012
CONFERENCE
SERIES



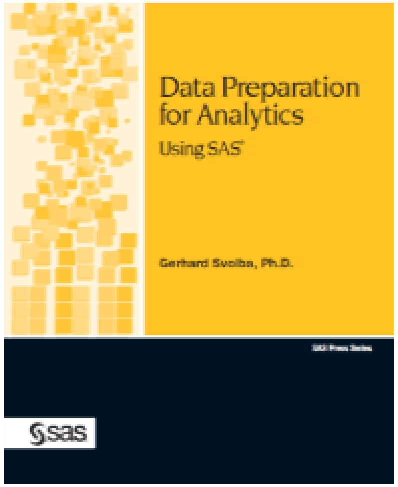
Data Quality for Analytics– and the consequences if it is not as good as you thought

Dr. Gerhard Svolba – SAS Austria
Las Vegas, October 8th, 2012

#analytics2012

„About the presenter"

- Product Manager for SAS Analytic Products
- Analytic Solution Architect at SAS Austria
- Author at SAS Press
- Enthusiastic Sailor



From this talk you can expect

- A practical and sportive introduction to „Data Quality for Analytics"
- The analytical viewpoint on data quality
- Answers to the questions (based on simulation studies)
  - How do missing values affect predictive power?
  - How much data do I need?

„A quick start to Sailing and Regattas“ (1)

Common start against the wind along a line between the start boat and a buoy

„A quick start to Sailing and Regattas“ (2)

Sailing „against“ the wind.

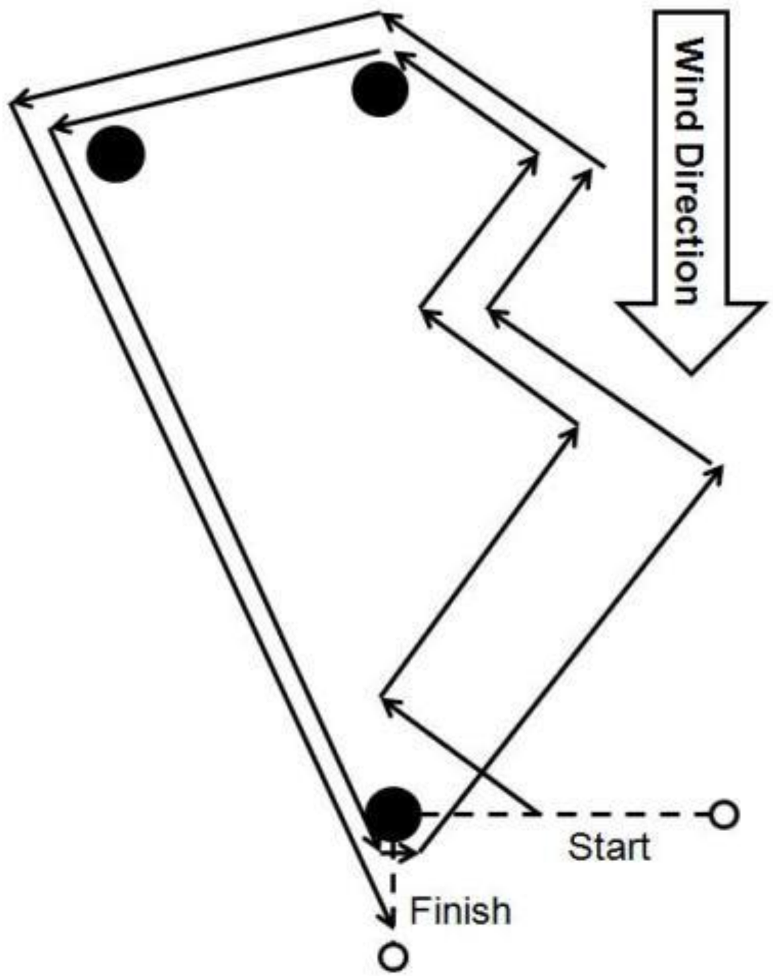
Tacking as the buoy cannot be reached on the direct way.

„A quick start to Sailing and Regattas“ (3)

After rounding the buoy, sailing with a spinnaker and wind from abaft to the next buoy.

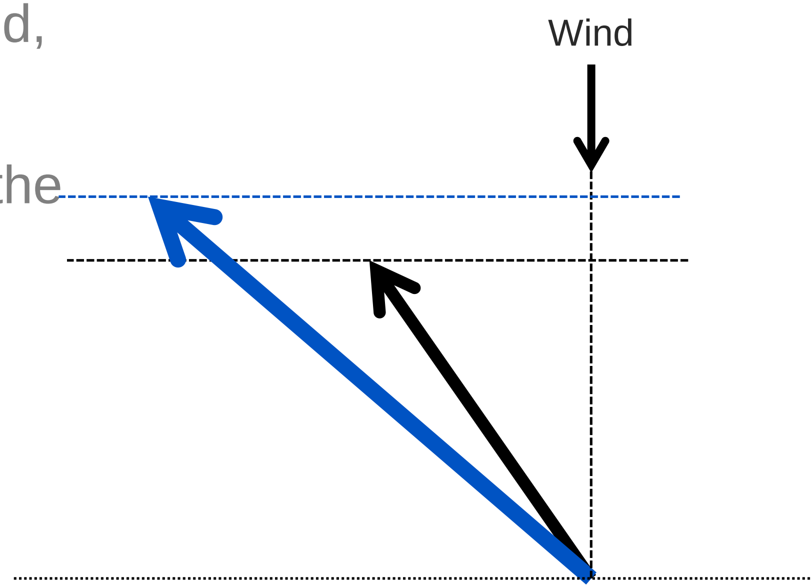
„A quick start to Sailing and Regattas“ (4)

Sketch of a regatta course with 3 buoys



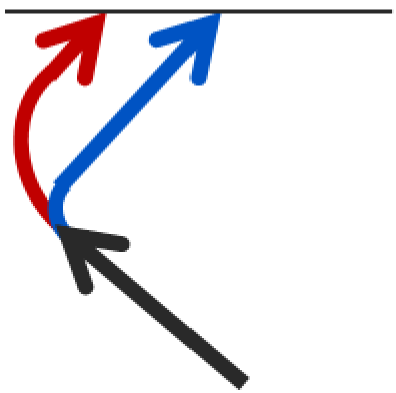
Here the statistician comes into play!
Analysis questions on optimizing sailing tactics (1)

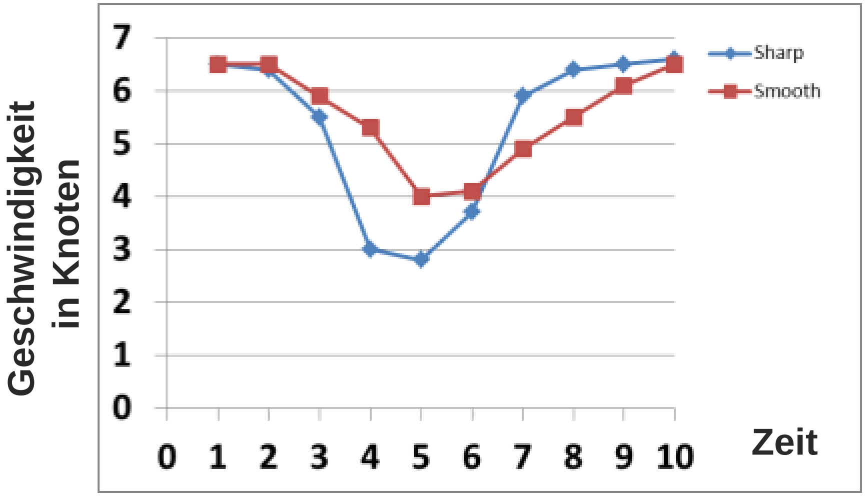
- Which course angle to the "true" wind shall be sailed?
  - The more acute the angle,
    - » the more direct the course,
    - » the shorter the distance to be sailed,
    - » but the lower the speed.
  - Depending on the wind strength and the sails



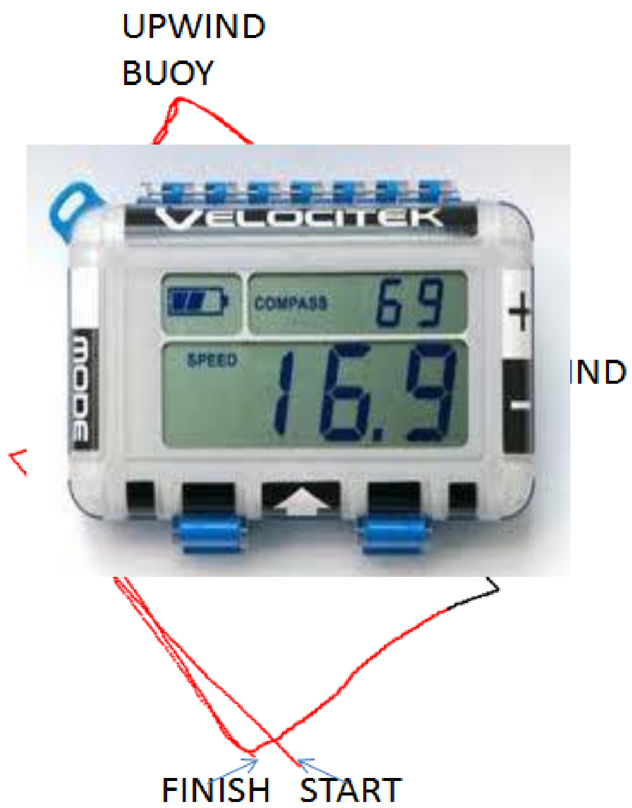
Here the statistician comes into play!
Analysis questions on optimizing sailing tactics (2)

- How shall the tacks be done?
  - **Rapidly, effective**: to quickly get to boat on a new course and gain speed?
  - **Round, flowing**: to make sure that the boat loses only little speed?





Available data for the sailing analyses (1): „GPS-Trackpoint Data“



- Longitude/Latitude Position
- Course (Compass heading)
- Speed

<MetadataTag name="SailorName" value="xxxx" />
</MetadataTags>
<CapturedTrack name="090521_131637" downloadedOn="2009-05-25T18:23:46.25+02:00" numberTrkpts="8680">
<MinLatitude>47.773464202880859</MinLatitude>
<MaxLatitude>47.804649353027344</MaxLatitude>
<MinLongitude>16.698064804077148</MinLongitude>
<MaxLongitude>16.74091911315918</MaxLongitude>
<DeviceInfo ftdiSerialNumber="VTQURQX9" />
<SailorInfo firstName="xxxx" lastName="yyyy" yachtClub="zzzz" />
<BoatInfo boatName="wwww" sailNumber="0000" boatClass="Unknown" hullNumber="0" />
<Trackpoints>
<Trackpoint dateTime="2009-05-21T13:49:24+02:00" heading="68.43" speed="5.906" latitude="47.792442321777344" longitude="16.727603912353516" />
<Trackpoint dateTime="2009-05-21T13:49:26+02:00" heading="59.38" speed="5.795" latitude="47.7924690246582" longitude="16.72768211364746" />
<Trackpoint dateTime="2009-05-21T13:49:28+02:00" heading="65.41" speed="6.524" latitude="47.792495727539062" longitude="16.727762222290039" />
<Trackpoint dateTime="2009-05-21T13:49:30+02:00" heading="62.2" speed="6.631" latitude="47.792518615722656" longitude="16.727849960327148" />
<Trackpoint dateTime="2009-05-21T13:49:32+02:00" heading="56.24" speed="6.551" latitude="47.792549133300781" longitude="16.727928161621094" />
<Trackpoint dateTime="2009-05-21T13:49:34+02:00" heading="60.56" speed="5.978" latitude="47.792579650878906" longitude="16.728004455566406" />
<Trackpoint dateTime="2009-05-21T13:49:36+02:00" heading="61.57" speed="7.003" latitude="47.792606353759766" longitude="16.728090286254883" />
<Trackpoint dateTime="2009-05-21T13:49:38+02:00" heading="52.03" speed="7.126" latitude="47.792636871337891" longitude="16.728176116943359" />

Available data for the sailing analyses (2): „Manual data collection“

- Composition of crew
- Sail size & type
- Wind speed and direction
- Placement in the race
- Other Comments

Datum	Regatta	Wettfahrt	Steuermann	Mittelmann	Spimann	Grossegel	Vorsegel	Spi	Spi gesetzt	Windstärke	Windrichtung
21.05.2009	Ruster Segeltage	1	Günter	Christian	Gerhard	Binder	2	Rot	1	3-4	S
21.05.2009	Ruster Segeltage	2	Günter	Christian	Gerhard	Binder	2	Rot	1	3-4	S
21.05.2009	Ruster Segeltage	3	Günter	Christian	Gerhard	Binder	2	Rot	1	3-4	S
22.05.2009	Ruster Segeltage	4	Günter	Christian	Gerhard	Binder	1	Rot	1	1-2	NW
23.05.2009	Ruster Segeltage	5	Günter	Christian	Gerhard	Binder	3	Rot	1	2-3	NW
23.05.2009	Ruster Segeltage	6	Günter	Christian	Gerhard	Binder	2	Rot	1	2-3	NW
23.05.2009	Ruster Segeltage	7	Günter	Christian	Gerhard	Binder	2	Rot	1	2-3	NW
20.06.2009	Blaues Band	1	Günter	Karl	Gerhard	Binder?	2	Rot	1	4-5	NW
27.06.2009	3 Insel	1	Günter	Karl	Gerhard	Binder	1	Rot	1	2-3	NW
27.06.2009	3 Insel	2	Günter	Karl	Gerhard	Binder	1	Rot	1	2	NW
27.06.2009	3 Insel	3	Günter	Karl	Gerhard	Binder	1	Rot	1	2	NW
28.06.2009	3 Insel	4	Günter	Karl	Gerhard	Binder	1	Rot	1	1	NW
28.06.2009	3 Insel	5	Günter	Karl	Gerhard	Binder	1	Rot	1	1	NW
25.07.2009	CBS-Cup	1	Günter	Karl	Gerhard	Binder	3	Rot	0	6	NW
26.07.2009	CBS-Cup	2	Günter	Karl	Gerhard	Binder	2	Rot	1	2-3	NW
26.07.2009	CBS-Cup	3	Günter	Karl	Gerhard	Binder	2	Rot	1	2-3	NW
26.07.2009	CBS-Cup	4	Günter	Karl	Gerhard	Binder	1	Rot	1	2	NW
19.09.2009	Absegeln	1	Günter	Michael Reite	Marlene + M	Binder	1	Rot	1	1	NW

Data quality issues in the case study and in business analysis are similar

- Failure of the GPS device because of low temperatures and bad batteries
- Trim settings of the boat were not documented
- Manual records: sometimes patchy, often only created after the event
- In rare cases: Long / Lat positioning delayed → miscalculation of speed
- Data transfer: GPS Device → (XML) → PC
  XML / Text → SAS
- Only 97 tacks documented in the data the first year

Data quality issues in the case study and in business analysis are similar (cont.)

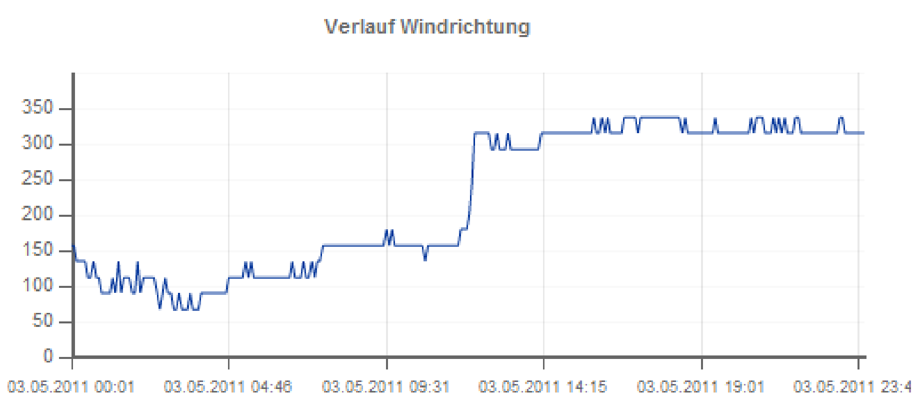
- Data Cleaning: data collection from turning on and turning of the device
- No GPS track point data of other sail boats available
- Wind speed and direction data are not collected on the boat
- External Data: Measuring station in the harbor. Different time intervals. No historical availability.

Availability and usability of data on the example of wind and weather data

▶ Ruster Bucht / Neusiedler See

▶ Wetterstation Rust 10.05.2011, 12:10 Uhr

Quelle: www.byc.at



Categorization of data quality issues

- Failure of the GPS device because of low temperatures and bad batteries
- Trim settings of the boat were not documented
- Manual records: sometimes patchy, often only created after the event

Data Completeness

- In rare cases: Long / Lat positioning delayed → miscalculation of speed
- Data transfer: GPS Device → (XML) → PC XML / Text → SAS

Data Correctness

- Only 97 tacks documented in the data the first year

Data Quantity

Categorization of data quality issues (cont.)

- Data Cleaning: data collection from turning on and turning of the device

Data Usability

- No GPS track point data of other sail boats available
- Wind speed and direction data are not collected on the boat
- External Data: Measuring station in the harbor. Different time intervals. No historical availability.

Data Availability

Typical criteria for data quality for analytics

- **Data Availability**
  - Actual data, historic data, historic snapshot of data

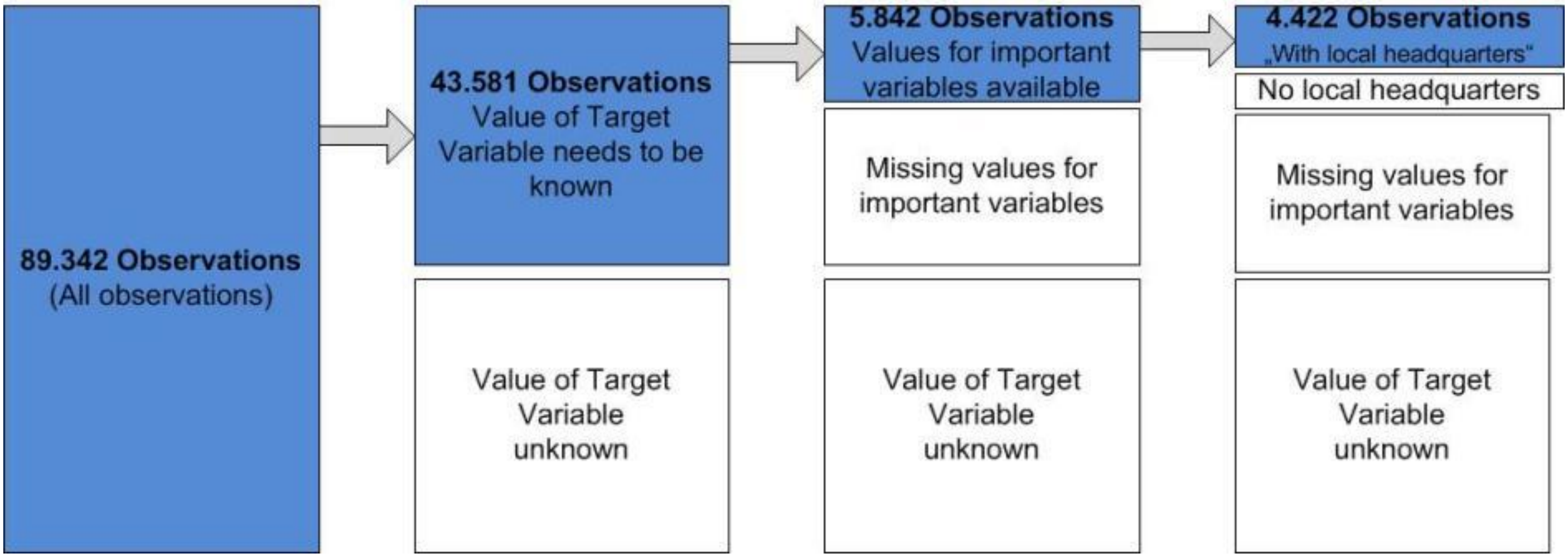
The term „Historic Data“ needs to be defined very precisely

	January					
	22	23	24	25	26	27
Rented Cars	**18.912**	**17.730**	**17.618**	**16.708**	**17.899**	**16.855**
Bookings (per day before)	18.853	17.729	17.616	16.510	17.728	16.843
Bookings (day -2)		17.693	17.617	16.512	17.727	16.881
Bookings (day -3)			17.701	16.511	17.678	16.709
Bookings (day -4)				16.666	17.675	16.707
Bookings (day -5)					17.619	16.513
Bookings (day -6)						16.509

Typical criteria for data quality for analytics

- **Data Availability**
  - Actual data, historic data, historic snapshot of data
  - Ensure periodic availability of data
  - Level of granularity: aggregations or detail data
- **Data Quantity**
  - Number of analysis subjects and events, length of observations period

The number of usable observations for the analysis reduces quickly

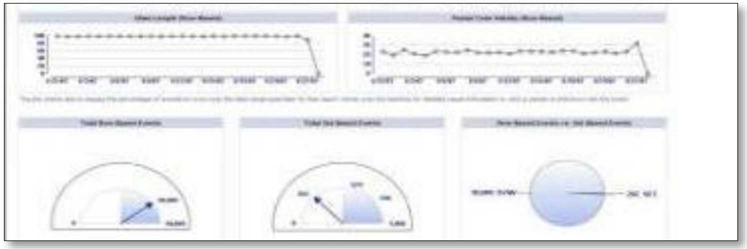


Typical criteria for data quality for analytics

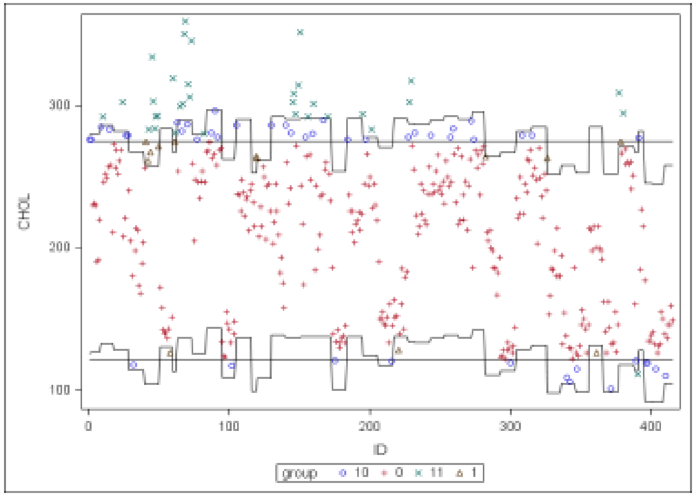
- **Data Availability**
  - Actual data, historic data, historic snapshot of data
  - Ensure periodic availability of data
  - Level of granularity: aggregations or detail data
- **Data Quantity**
  - Number of analysis subjects and events, length of observations period
- **Data Completeness**
  - Random or systematic missing values, patterns
  - Effort to get complete data
- **Data Correctness**
  - Univariate and multivariate plausibility checks
- **Statistical Features**
  - Correlation, variability, distributions

SAS helps to PROFILE data quality

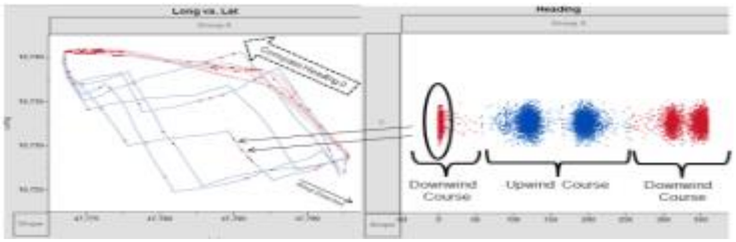
- DataFlux® / SAS Data Management Platform, Base SAS



- SAS® Enterprise Miner,
  SAS® STAT, SAS® ETS
  SAS® Forecast Server
  - Complex patterns of missing values
  - Outliers detection based on multivariate methods
  - Early detection of predictive power and variable importance



- JMP® for interactive visual data quality control



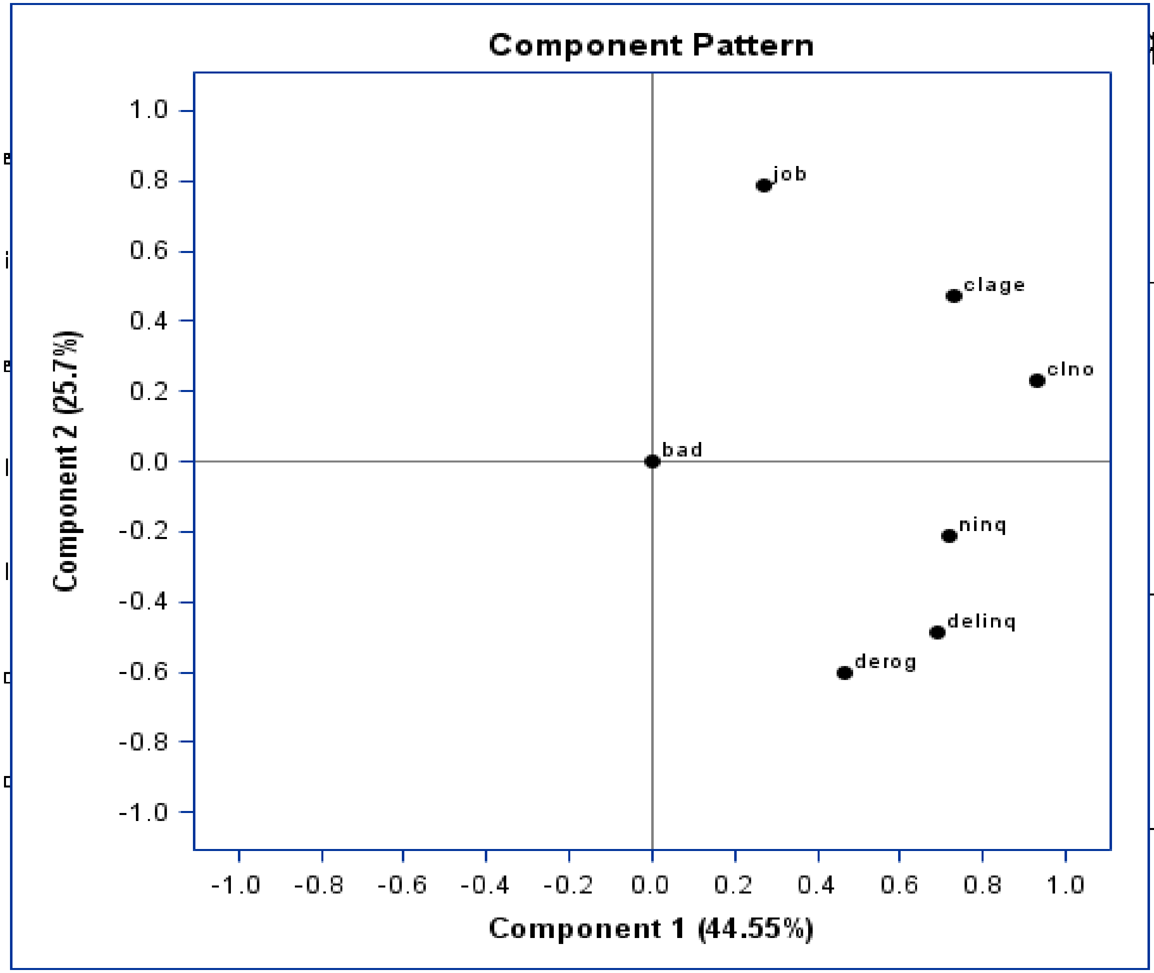
Profiling the pattern of missing values

(macros can be downloaded from www.sascommunity.org)

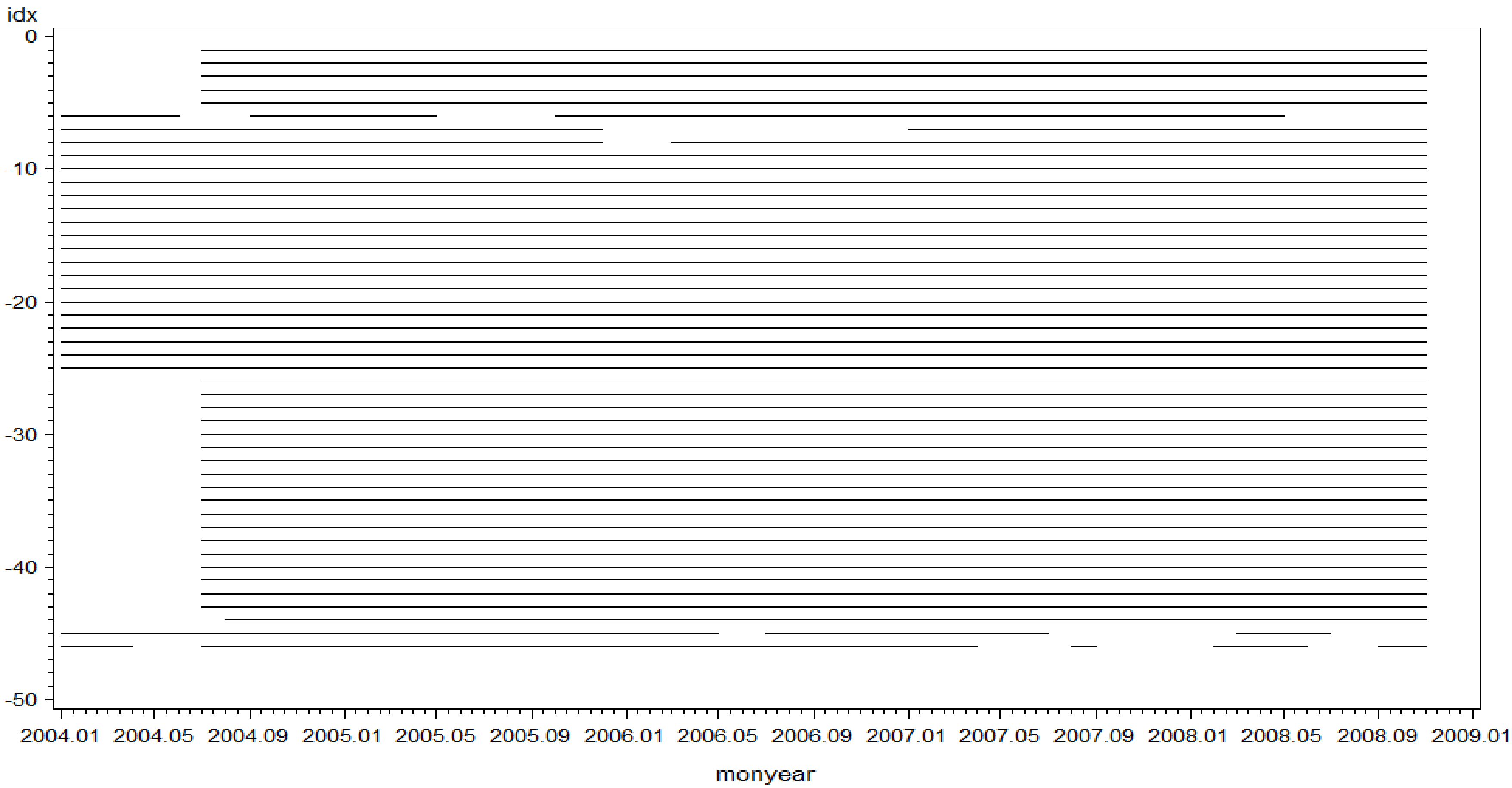
Descriptive Pattern

Variable Clustering

Principal Components



Profiling the structure of missing values and zero values in time series data
(macros can be downloaded from www.sascommunity.org)

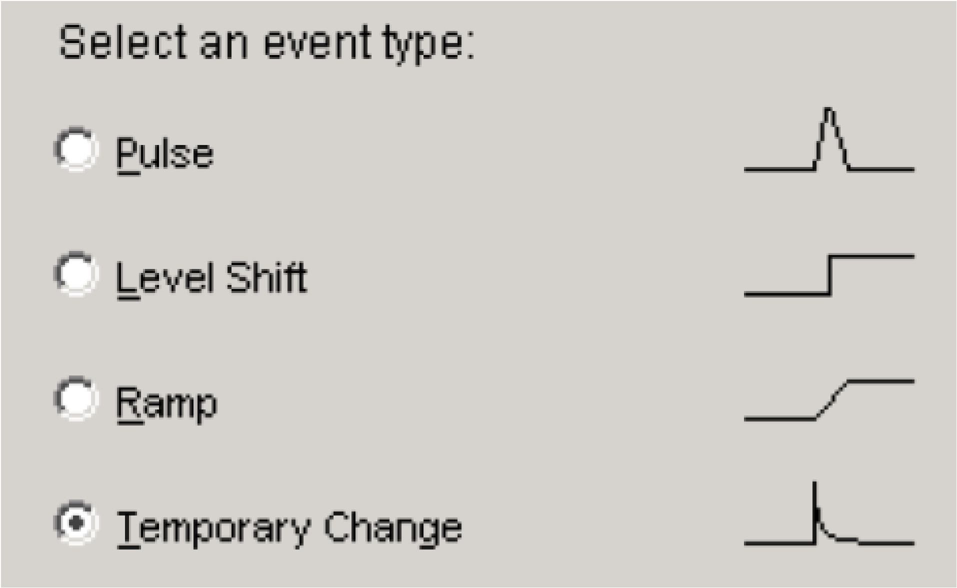


Analytics2012 CONFERENCE SERIES

#analytics2012

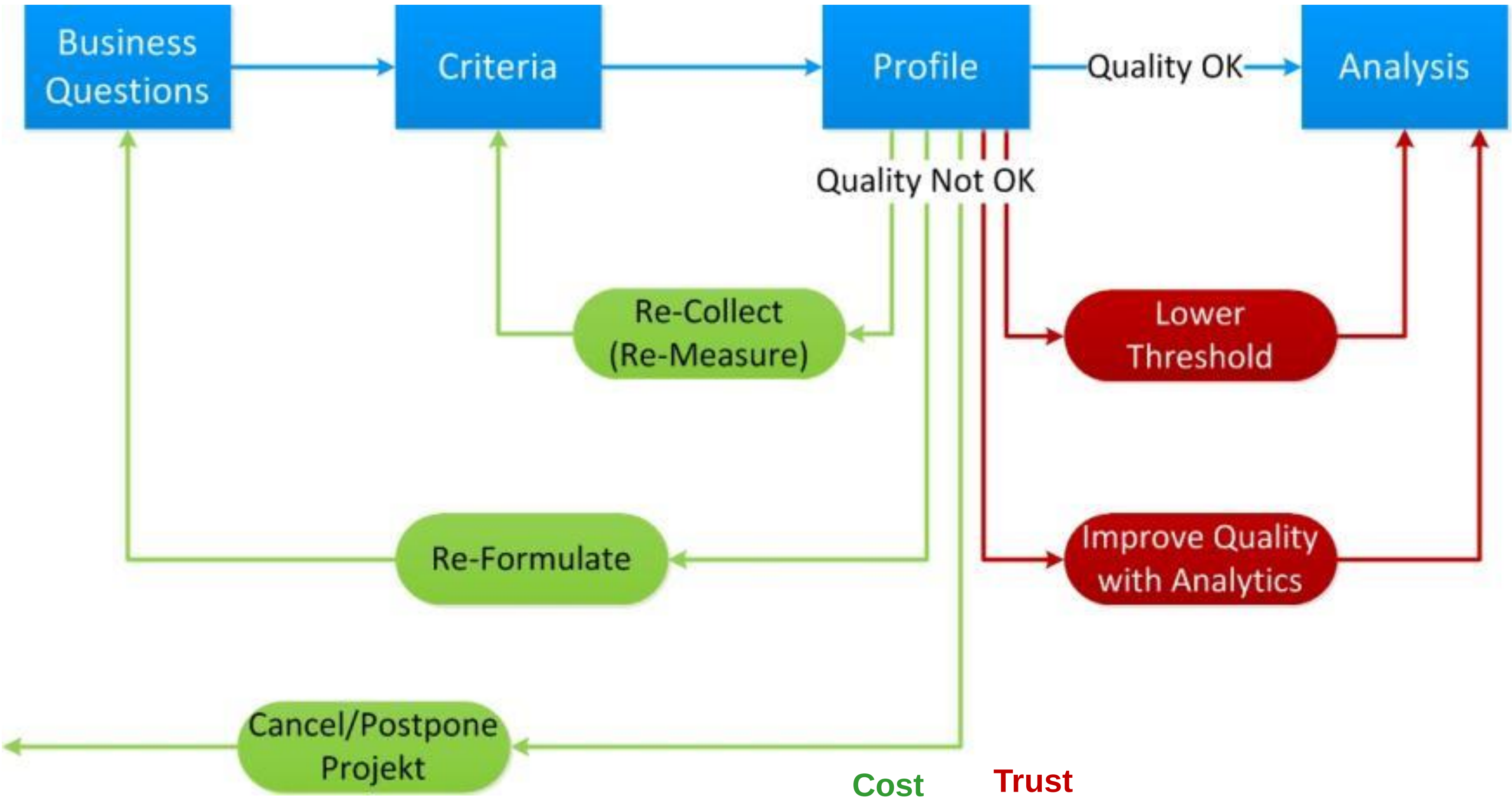
SAS helps to IMPROVE data quality

- Imputation of missing values
- Calculation of individual replacement values
- Treat „exceptional“ subgroups and time periods differently in the model



- Similarity measures for standardization and record matching
- Methods for rare events
- Sample size planning

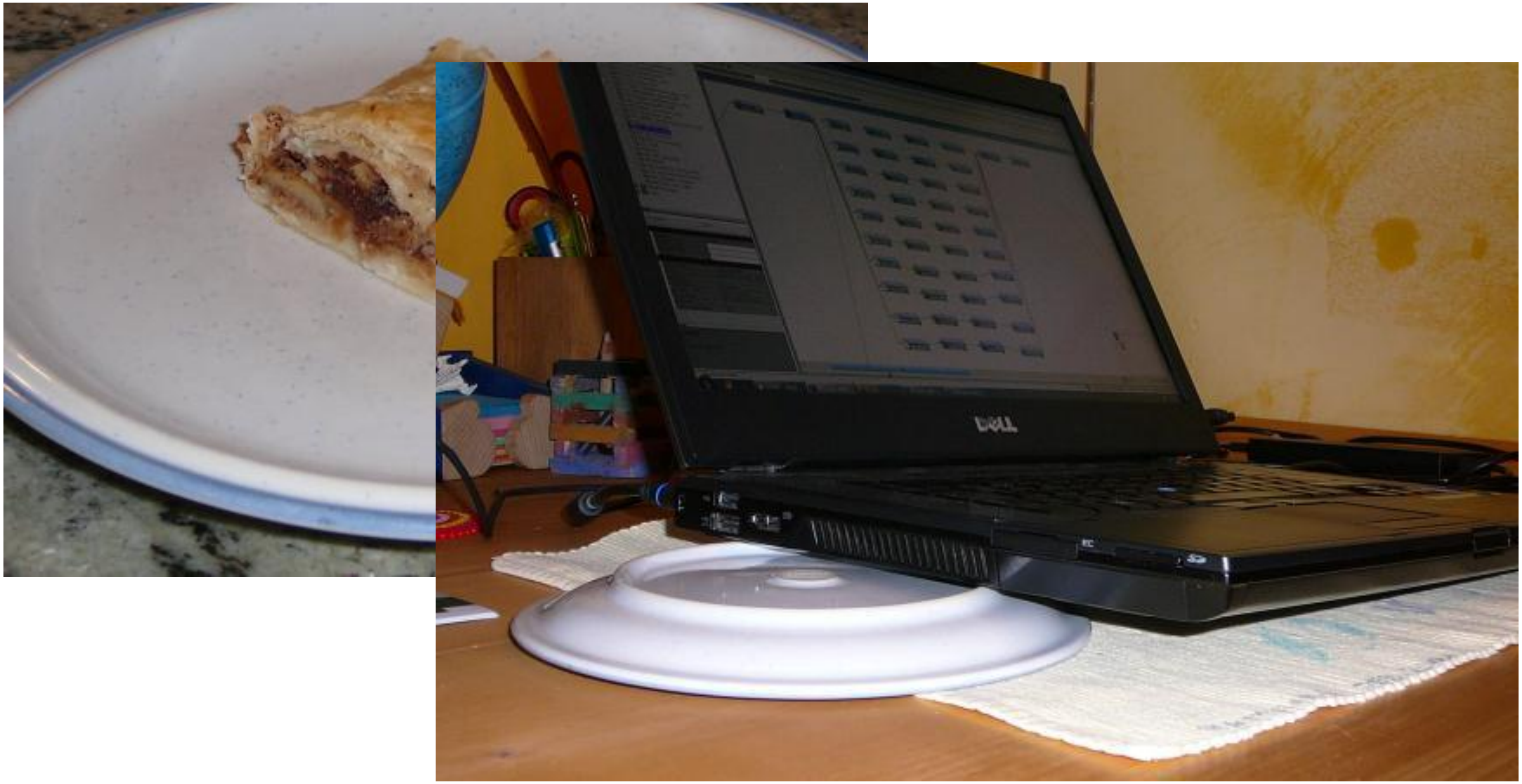
These are your options, if you learn that data quality is poor



Cost
Time, Delays
No Results

Trust
Risk of wrong decisions
Insignificance

Simulation studies for the consequences of poor data quality



The consequences of the following effects have been studied

- How do missing values affect predictive power?
  - Random and systematic missing values (SIM_1)
- How much data do I need?
  - Varying the available number of observations and events (SIM_2)
  - Gradually increasing the available length of data history (SIM_3)

- Other questions / simulations
  - Withholding the set of the most important variables
  - Introducing random and systematic bias in the input and target variables in predictive modeling
  - Effect of random and systematic missing values and bias in time series forecasting

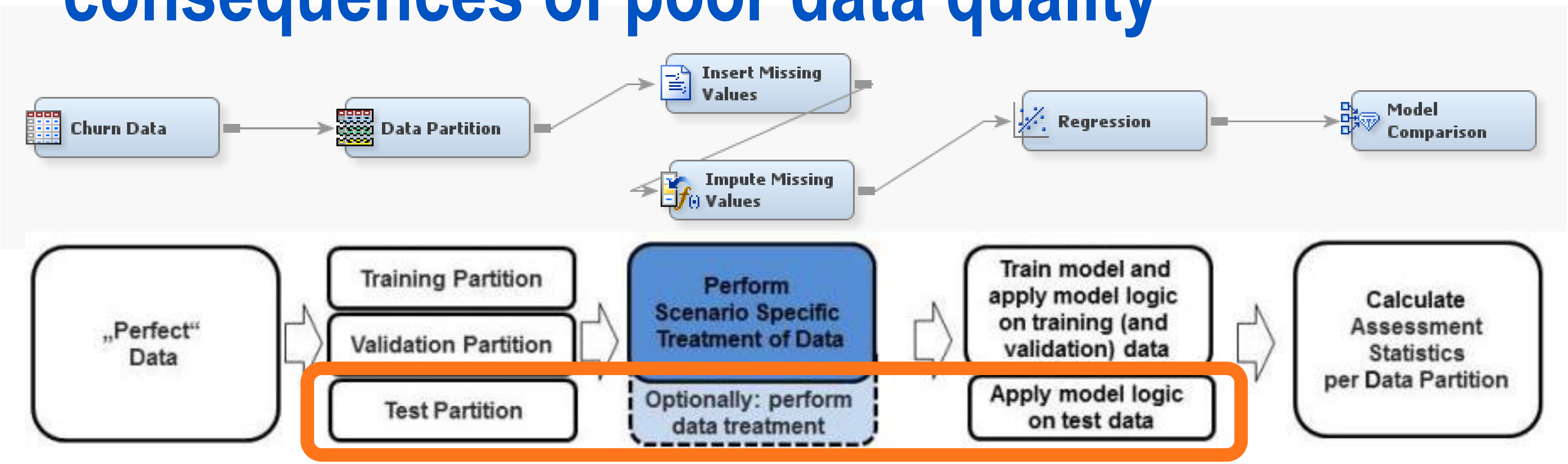
Real life data is used for the simulation studies

- Four real life datasets from different industries with a binary target variable were used
- Drop variable with > 5 % of missing values
- Drop observations, if ≤ 5 % of missing values
- Run multiple model cycles to retrieve a stable model with good predictive power
- Freeze the list of variables for the simulations

ID	TRAVTIM	BLUEBOO	INITDAT	RED_C	AGE	INCOME	TDAT	CAR_TYP	RED_C	AGE	HOMEKID	YOJ	INCOME	KID	YOJ	INCOME
13276104	22	$14.940	29Jul1998	yes	43	$91.449	Jul1998	Sedan	yes	43	0	11	$91.449	0	11	$91.449
8568865	48	$18.510	#######	no	50	$106.952	######	Van	no	50	0	7	$106.952			
3764824	25	$17.600	02Oct1988	yes	55	$59.162	Oct1988	Van	yes	55	0	11	$59.162	0	7	$106.952
6916555	7	$25.660	30Jan1995	no	46	$59.162	Jan1995	Panel Truck	no	46	0	14	$59.162	0	11	$59.162
2693415	12	$25.810	21Jan1991	no	41	$92.842	Jan1991	Panel Truck	no	41	0	7	$92.842	0	14	$59.162
3969722	38	$16.640	12May1993	yes	39	$35.081	May1993	Van	yes	39	0	10	$35.081			
1737378	38	$29.450	27Sep1989	no	43	$145.353	Sep1989	Van	no	43	2	17	$145.353	0	7	$92.842
7082153	48	$9.280	02Jan1994	no	52	$0	Jan1994	SUV	no	52	0	0	$0	0	10	$35.081
2589510	22	$6.450	24Nov1998	no	35	$14.508	Nov1998	Pickup	no	35	0	7	$14.508	2	17	$145.353
2556136	5	$29.270	23Nov1993	yes	40	$57.474	Nov1993	Panel Truck	yes	40	0	11	$57.474	0	0	$0
7530371	62	$4.600	09Nov1989	yes	31	$26.520	Nov1989	Sedan	yes	31	1	12	$26.520	0	7	$14.508
2130575	24	$23.450	22Sep1994	no	42	$52.988	Sep1994	Sedan	no	42	0	12	$52.988	0	11	$57.474
4012002	14	$28.560	12May1986	yes	48	$52.988	May1986	Panel Truck	yes	48	0	12	$52.988	1	12	$26.520
9350798	31	$33.710	02Jul1991	no	49	$52.988	2Jul1991	Panel Truck	no	49	0	9	$52.988			
2990245	22	$33.320	22Feb1990	yes	46	$52.988	Feb1990	Panel Truck	yes	46	0	14	$52.988	0	12	$52.988
3939254	30	$10.910	#######	yes	45	$61.931	######	Pickup	yes	45	0	11	$61.931	0	12	$52.988
5366048	40	$23.230	06Jun1992	yes	48	$61.509	Jun1992	Panel Truck	yes	48	0	9	$61.509	0	9	$52.988
8033252	35	$22.050	06Jul1985	no	44	$139.330	Jul1985	Sports Car	no	44	2	15	$139.330	0	14	$52.988
4660671	41	$17.470	15Jun1987	no	38	$0	Jun1987	Van	no	38	0	0	$0	0	11	$61.931
5935828	18	$23.390	07Jul1980	yes	42	$76.226	Jul1980	Panel Truck	yes	42	0	9	$76.226	0	9	$61.509
1704832	36	$15.120	05Jun1997	no	52	$68.992	Jun1997	Sedan	no	52	0	9	$68.992	2	15	$139.330
42131636	8	$30.150	#######	no	43	$132.561	######	Panel Truck	no	43	0	11	$132.561	0	0	$0
3707484	21	$7.500	29Jul1985	yes	60	$125.893	Jul1985	Pickup	yes	60	0	9	$125.893			
7182942	14	$17.550	21Oct1995	no	37	$123.520	Oct1995	Van	no	37	0	12	$123.520	0	9	$76.226
5388757	33	$29.210	02Jun1991	yes	47	$45.257	Jun1991	Panel Truck	yes	47	0	13	$45.257	0	9	$68.992
5209593	53	$13.050	12Dec1993	no	40	$75.516	Dec1993	Sports Car	no	40	3	9	$75.516	0	11	$132.561
5684737	40	$36.120	16Dec1990	yes	47	$104.271	Dec1990	Panel Truck	yes	47	0	12	$104.271	0	9	$125.893
4538673	35	$28.180	#######	no	33	$111.427	######	Van	no	33	1	12	$111.427	0	12	$123.520
5820861	11	$17.300	#######	yes	50	$111.427	######	Van	yes	50	0	14	$111.427	0	13	$45.257
6804259	32	$11.620	30Mar1995	no	51	$50.166	Mar1995	Pickup	no	51	0	9	$50.166	3	9	$75.516
93412915	50	$14.530	09Dec1982	no	43	$48.184	Dec1982	Sports Car	no	43	3	14	$48.184	0	12	$104.271
46157391	24	$21.990	21Jun1983	no	49	$22.059	Jun1983	Pickup	no	49	0	8	$22.059	1	12	$111.427
2252155	35	$12.180	#######	yes	34	$23.571	######	Pickup	yes	34	1	10	$23.571			
6833784	45	$27.890	05Nov1989	no	55	$55.409	Nov1989	Panel Truck	no	55	0	14	$55.409	0	14	$111.427
5039064	48	$8.460	05Sep1987	yes	48	$39.613	Sep1987	Pickup	yes	48	0	10	$39.613	0	9	$50.166
19707619	9	$34.510	07Jun1988	no	45	$23.773	Jun1988	Van	no	45	3	15	$23.773	3	14	$48.184
3457713	17	$31.390	25Apr1996	yes	41	$55.364	Apr1996	Panel Truck	yes	41	0	8	$55.364	0	8	$22.059
8308556	44	$23.320	01Sep1983	no	55	$163.158	Sep1983	Panel Truck	no	55	0	16	$163.158	1	10	$23.571
6429873	59	$10.350	21Jan1982	yes	52	$24.590	Jan1982	Sedan	yes	52	0	12	$24.590	0	14	$55.409
9309292	45	$27.020	04Apr1992	no	54	$107.808	Apr1992	Pickup	no	54	0	15	$107.808	0	10	$39.613
39176001	42	$7.470	22Dec1987	no	39	$59.685	Dec1987	Pickup	no	39	2	12	$59.685	3	15	$23.773
8419408	51	$6.720	26Jun1991	no	44	$146.267	Jun1991	Pickup	no	44	0	4	$146.267	0	8	$55.364
60189132	10	$6.120	23Apr1986	no	45	$1.158	Apr1986	SUV	no	45	3	4	$1.158	0	16	$163.158
7921960	32	$13.160	23Oct1984	no	43	$1.158	Oct1984	Sedan	no	43	0	11	$1.158	0	12	$24.590
														0	15	$107.808

7921960	07Oct1997	32	Commercia	2577125	$13.160	23Oct1984	Sedan	no	43	0	11	$1.158	2	12	$59.685
8419408	25Jun1995	51	Commercia	2465546	$6.720	26Jun1991	Pickup	no	44				0	4	$146.267
60189132	20Apr1996	10	Commercia	2546478	$6.120	23Apr1986	SUV	no	45				3	4	$1.158
7921960	07Oct1997	32	Commercia	2577125	$13.160	23Oct1984	Sedan	no	43				0	11	$1.158

Simulation studies help to quantify the consequences of poor data quality



„Untouched“ Testdata act as „Scoring Data“

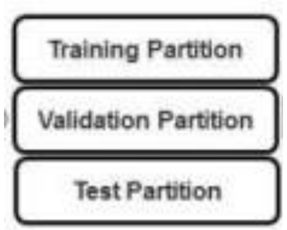
- Delete Observations
- Insert Missings
- Replace Missings

Use frozen list of variables on scenario data

Process for the missing value scenarios (SIM_1)

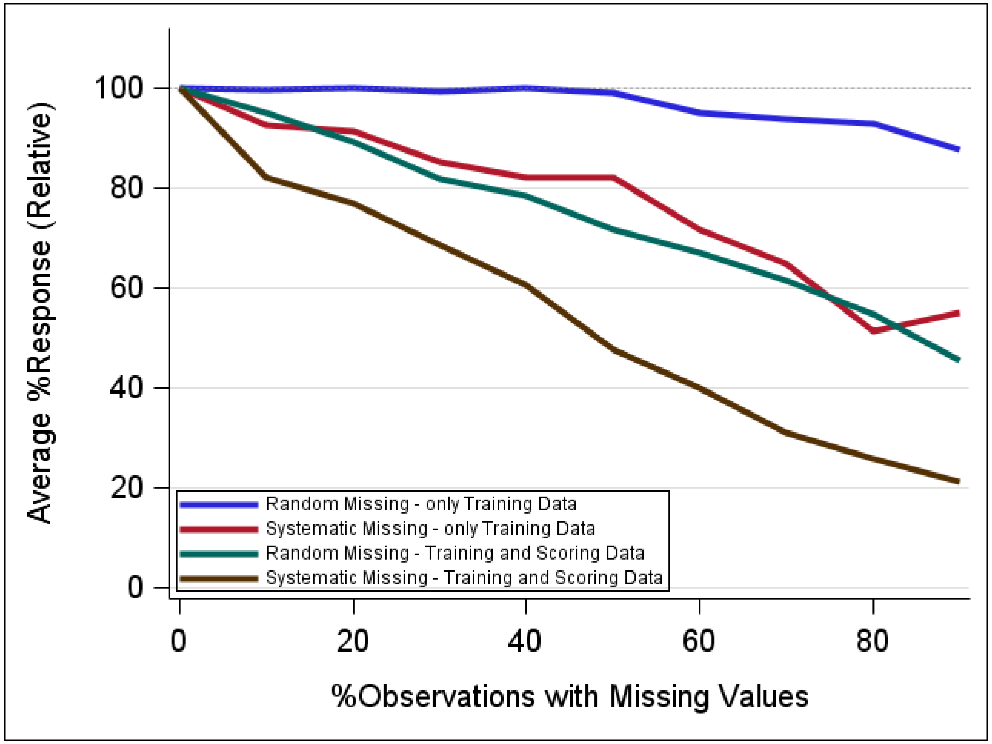
- For a specified proportion of observations (10%, …)
  - Set interval and nominal input variables to missing
    - » Random selection
    - » Systematic selection based on segments
  - Impute missing values with the IMPUTE node of SAS Enterprise Miner
- Train the model with frozen set of variables
- Optionally: Perform „treatment“ also for scoring data
- Assess model quality

ID	TRAVTIM	BLUEBOO	INITDAT	RED_C	AGE	INCOME
13276104	22	$14.940	29Jul1998	yes	43	$91.449
8568865	48	$18.510	#######	no	40	$106.952
3764824	25	$17.600	02Oct1988	yes	55	$59.162
6916555	7	$17.230	30Jan1995	no	46	$59.162
2693415	12	$25.810	21Jan1991	no	40	$59.162
3969722	38	$16.640	12May1993	yes	39	$35.081
1737378	38	$29.450	27Sep1989	no	43	$145.353
7082153	48	$9.280	02Jan1994	no	52	$0
2589510	22	$6.450	24Nov1998	no	35	$14.508
2556136	23	$29.270	23Nov1993	yes	40	$57.474
7530371	62	$4.600	09Nov1989	yes	31	$26.520
2130575	24	$23.450	22Sep1994	no	42	$52.988
4012002	14	$28.560	12May1986	yes	48	$52.988
9350798	31	$33.710	02Jul1991	no	49	$52.988
2990245	23	$33.320	22Feb1990	yes	40	$52.988
3939254	30	$10.910	#######	yes	45	$61.931
5366048	40	$23.230	06Jun1992	yes	48	$61.509
8033252	35	$22.050		no	44	$139.330
4660671	41	$17.470	15Jun1987	no	38	$0
5935828	18	$23.390	07Jul1980	yes	42	$76.226
1704832	36	$17.230	05Jun1997	no	52	$59.162
42131636	8	$30.150	#######	no	43	$132.561
3707484	21	$7.500	29Jul1985	yes	40	$125.893
7182942	14	$17.550	21Oct1995	no	40	$123.520
5388757	33	$29.210	02Jun1991	yes	47	$45.257
5209593	53	$13.050	12Dec1993	no	40	$75.516
5684737	40	$36.120	16Dec1990	yes	47	$104.271
4538673	35	$28.180	#######	no	33	$111.427
5820861	11	$17.300	#######	yes	50	$111.427
6804259	23	$11.620	30Mar1995	no	51	$50.166
93412915	50	$14.530	09Dec1982	no	43	$48.184
46157391	24	$21.990	21Jun1983	no	40	$22.059
2252155	35	$12.180	#######	yes	40	$23.571
6833784	45	$27.890	05Nov1989	no	55	$55.409
5039064	48	$8.460	05Sep1987	yes	48	$39.613
19707619	9	$17.230	07Jun1988	no	45	$23.773
3457713	17	$31.390	25Apr1996	yes	40	$55.364
8308556	44	$23.320	23Jan1900	no	55	$163.158
6429873	23	$10.350	21Jan1982	yes	52	$59.162
9309292	45	$27.020	04Apr1992	no	54	$107.808
39176001	42	$7.470	22Dec1987	no	39	$59.685
8419408	51	$6.720	26Jun1991	no	44	$146.267
60189132	10	$6.120	23Apr1986	no	45	$1.158
7921960	32	$13.160	23Oct1984	no	43	$1.158



Findings of the missing value scenarios

- Random missing values in training data only have limited effect.
- Missing values in the scoring data as well affect much more.
- Systematic missing values have a much larger effect.
- Things that matter:
  - Not only the proportion of missing values but especially the type
  - Missing values in the scoring data



Analytics2012 CONFERENCE SERIES

Quantifying the results of the missing value scenarios

- Running a general linear model:

Response = f(%missing, Systematic_YN, ScoringData_YN)

Parameter	Value	Interpretation
Intercept	19.29	Response with no missing values
%missing	- 0.1	10 % missing ~ 1% less response
Systematic_YN	- 3.6	Systematic error causes 3.6 % less response
Scoring_YN	- 4.23	Missings in scoring data cause 4.23 % less response

Studying the effect of data quantity in event prediction (SIM_2)

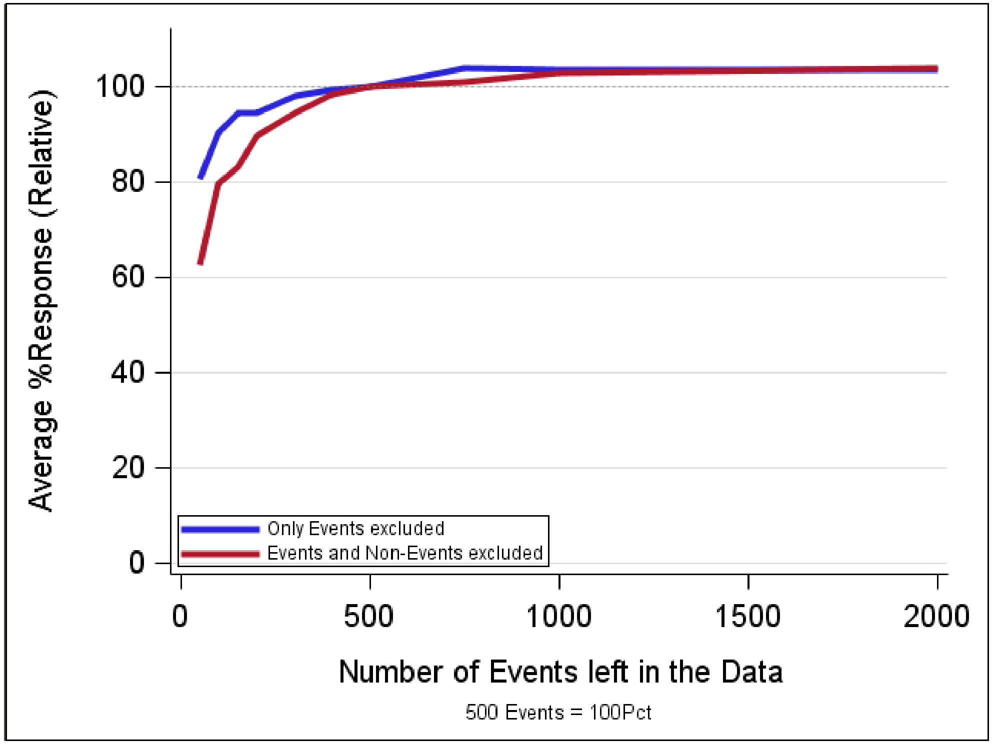
- Randomly selected observations and events were deleted from the data
- Additional observations and events provide increase in % Correct Response rate
- But:
  - Linear or non-linear effect
  - How can this effect be quantified?
  - Do also non-events contribute to an increase?
  - Is it worth waiting for more events?

ID	TRAVTIM	BLUEBOO	INITDAT	RED_C	AGE	INCOME
13276104	22	$14.940	29Jul1998	yes	43	$91.449
8568865	48	$18.510	#######	no	50	$106.952
3764824	25	$17.600	02Oct1988	yes	55	$59.162
2693415	12	$25.810	21Jan1991	no	41	$92.842
3969722	38	$16.640	12May1993	yes	39	$35.081
1737378	38	$29.450	27Sep1989	no	43	$145.353
7082153	48	$9.280	02Jan1994	no	52	$0
2556136	5	$29.270	23Nov1993	yes	40	$57.474
7530371	62	$4.600	09Nov1989	yes	31	$26.520
2130575	24	$23.450	22Sep1994	no	42	$52.988
4012002	14	$28.560	12May1986	yes	48	$52.988
9350798	31	$33.710	02Jul1991	no	49	$52.988
2990245	22	$33.320	22Feb1990	yes	46	$52.988
3939254	30	$10.910	#######	yes	45	$61.931
5935828	18	$23.390	07Jul1980	yes	42	$76.226
1704832	36	$15.120	05Jun1997	no	52	$68.992
42131636	8	$30.150	#######	no	43	$132.561
3707484	21	$7.500	29Jul1985	yes	60	$125.893
7182942	14	$17.550	21Oct1995	no	37	$123.520
5388757	33	$29.210	02Jun1991	yes	47	$45.257
5209593	53	$13.050	12Dec1993	no	40	$75.516
5684737	40	$36.120	16Dec1990	yes	47	$104.271
5820861	11	$17.300	#######	yes	50	$111.427
6804259	32	$11.620	30Mar1995	no	51	$50.166
93412915	50	$14.530	09Dec1982	no	43	$48.184
46157391	24	$21.990	21Jun1983	no	49	$22.059
2252155	35	$12.180	#######	yes	34	$23.571
6833784	45	$27.890	05Nov1989	no	55	$55.409
5039064	48	$8.460	05Sep1987	yes	48	$39.613
19707619	9	$34.510	07Jun1988	no	45	$23.773
8308556	44	$23.320	01Sep1983	no	55	$163.158
6429873	59	$10.350	21Jan1982	yes	52	$24.590
9309292	45	$27.020	04Apr1992	no	54	$107.808
39176001	42	$7.470	22Dec1987	no	39	$59.685
60189132	10	$6.120	23Apr1986	no	45	$1.158
7921960	32	$13.160	23Oct1984	no	43	$1.158
3457713	17	$31.390	25Apr1996	yes	41	$55.364
8308556	44	$23.320	01Sep1983	no	55	$163.158
6429873	59	$10.350	21Jan1982	yes	52	$24.590
9309292	45	$27.020	04Apr1992	no	54	$107.808
39176001	42	$7.470	22Dec1987	no	39	$59.685
8419408	51	$6.720	26Jun1991	no	44	$146.267
60189132	10	$6.120	23Apr1986	no	45	$1.158
7921960	32	$13.160	23Oct1984	no	43	$1.158

Findings of the data quantity scenarios

- Marginal benefits flattens out in the area of 500 to 1000 events
- Also non-events provide additional information especially in the area of up to 500 events

Varying the number of events and non-events

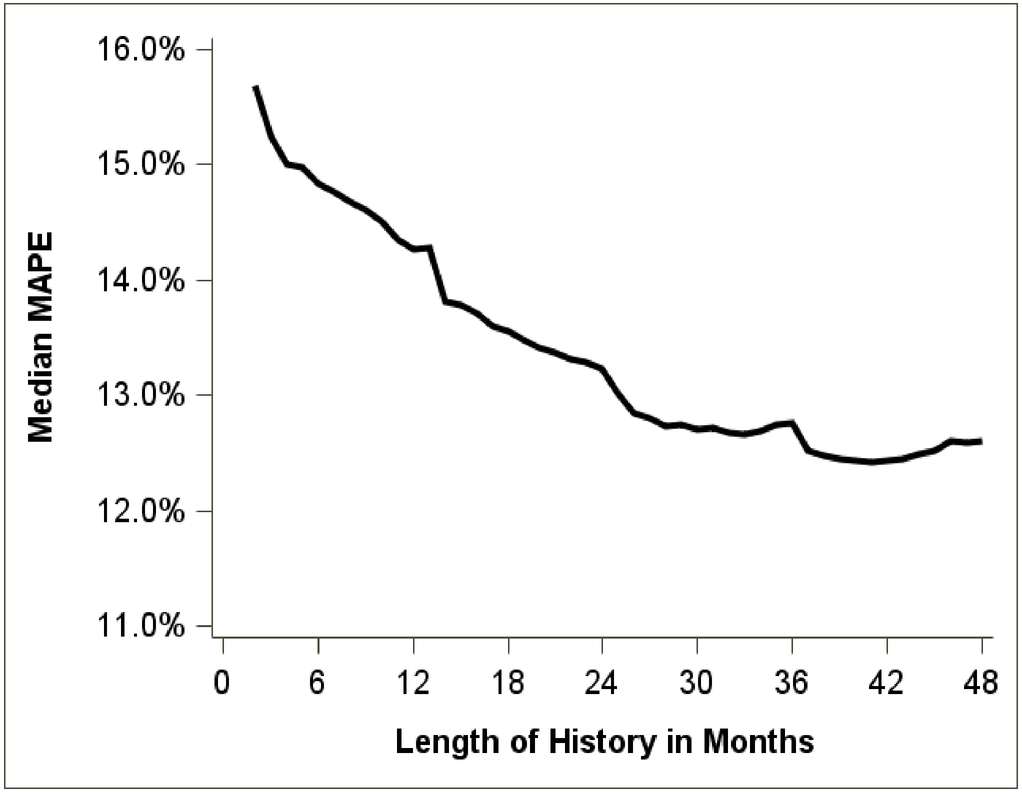


Gradually increasing the available length of data history in time series forecasting

- Business Questions
  - Is it possible to start time series analysis if only 18 months of history are available?
  - Do we have to wait for an additional history month?
  - What is the benefit of additional data management effort?
  - What is the best length of data history for time series forecasting?
- Methods
  - Simulation environment built with SAS High Performance Forecasting
  - 788 time series on monthly data from different industries
  - Minimum history for each time series: 48 months
  - Restricted to forecasting method „exponential smoothing“
  - Validation based on MAPE calculated on 12 lead months
  - Iterating by shifting the „zero-time“ over 12 months for better generalizability

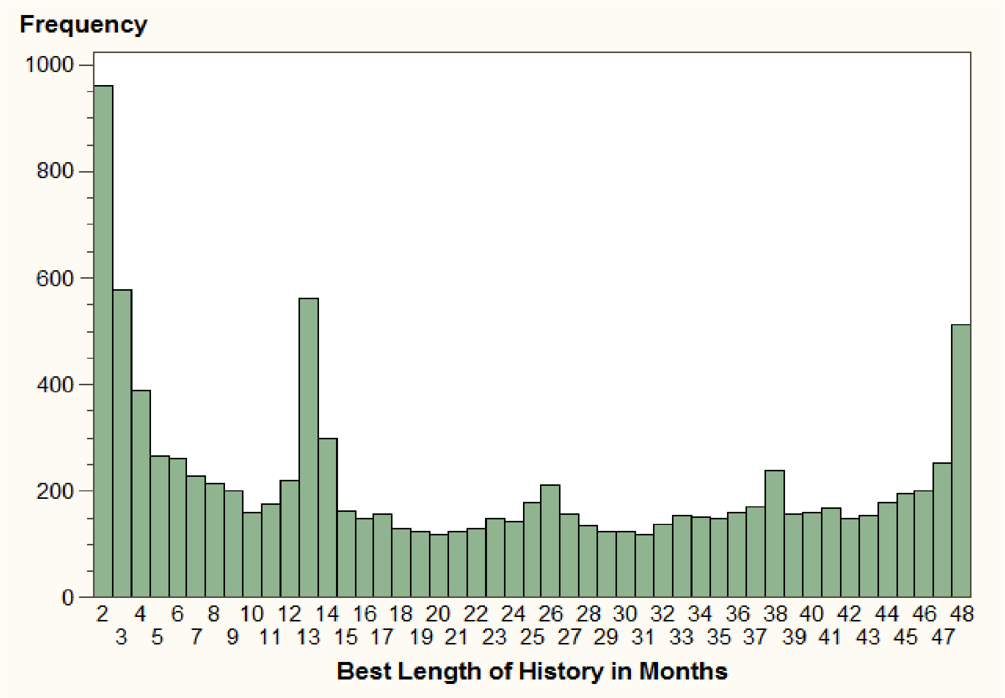
How far should we remember back?

- The expected decrease in MAPE with increasing history length can be seen.
- There is an exponential decrease in the additional value of additional months
- Larger steps after 12, 24 and 36 months can be seen.



What is the best length of data history for time series forecasting?

- Method: for each time series query how many history months give the smallest error for the future 12 months
- Results: Not in all cases it is beneficial to use a long data history.



Final takeaways

- Data Quality for Analytics is more
  - More requirements
  - More possibilities
- Get into details!
  - Random or systematic bias?
  - Permanent or historic/temporary problem?
- Quantity matters!
  - But balance effort and benefit!
- SAS helps to
  - Profile, Improve, Assess, Simulate

Data Quality for Analytics Using SAS

SAS Press, April 2012

Dr. Gerhard Svolba – sastools.by.gerhard@gmx.net – LinkedIn

http://www.sascommunity.org/wiki/Data_Quality_for_Analytics

- Analytics has additional requirements on data quality
- Analytics contributes methods for better data quality
- Simulation studies show the consequences of poor data quality on model quality